RU-EVAL-2019: EVALUATING ANAPHORA AND COREFERENCE RESOLUTION FOR RUSSIAN

Budnikov E. A.* (egor.budnikov@abbyy.com), Toldova S. Yu.* (toldova@yandex.ru), Zvereva D. S.*** (zverevads@gmail.com), Maximova D. M. ** (daria.maximova.m@gmail.com), Ionov M. I.**** (max.ionov@gmail.com)

* ABBYY, Moscow, Russia

** National Research University Higher School of Economics, Moscow, Russia

*** Moscow Institute of Physics and Technology, Moscow, Russia

**** Goethe University Frankfurt, Frankfurt am Main, Russia

RU-EVAL-2019: РАЗРЕШЕНИЕ АНАФОРЫ И КОРЕФЕРЕНТНОСТИ НА РУССКОМ ЯЗЫКЕ

Будников Е. А.* (egor.budnikov@abbyy.com), Толдова С. Ю.* (toldova@yandex.ru), Зверева Д. С.*** (zverevads@gmail.com), Максимова Д. М. ** (daria.maximova.m@gmail.com), Ионов М. И.**** (max.ionov@gmail.com)

* ABBYY, Москва, Россия

** Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), Москва, Россия

*** Московский Физико-Технический Институт, Москва, Россия

**** Goethe University Frankfurt, Франкфурт, Россия

This paper reports on one of the events within the evaluation campaign RU-EVAL-2019. The NLP task behind this even concerns anaphora and coreference resolution in Russian. The first evaluation event devoted to this issue was held in 2014. However, since then the NLP technologies and resources for Russian has changed greatly. The present event is organized in order to estimate the current state-of-the-art regarding this task and to compare various methods and principles implemented for Russian. In this paper, we define anaphora and coreference resolution tasks, describe the dataset with special attention to the annotation rules, clarify the evaluation procedure and discuss the results. Besides, we present a new open dataset with coreference annotations for Russian (523 documents, 5.7k chains with 25k mentions) that can be used for training and evaluation while developing new systems.

Keywords: coreference resolution, anaphora resolution, discourse processing, resource description

1. *Introduction*

Coreference resolution is the task of determining which mentions in a text refer to the same entity. Two mentions (i.e. textual phrases) are called coreferent if they refer to the same real-world objects or events. Several coreferent mentions are sometimes called a coreference chain.

(1) Theresa Mary May is a British politician serving as the current Prime Minister of the United Kingdom. She identifies herself as a one-nation conservative.

In (1) mentions "Theresa Mary May" and "She" are coreferent as they refer to the same person.

Coreference is a complex and multifaceted phenomenon with numerous factors in play and currently there were no reported results on methods or instruments for obtaining full and high quality solution for its resolution for Russian.

At the same time, the task is vital for a large number of the high-level NLP tasks, e.g.:

- Information Extraction
- Information Retrieval
- Sentiment Analysis (Opinion Mining)
- Question-answering system
- Machine translation
- Automatic Text Summarization

Consider examples (2) and (3):

(2) Peter$_1$ asked John to come home because he$_1$ was scared of being alone.

(3) Peter asked John$_1$ to come home because he$_1$ was working late again.

To answer the questions «Who of the boys feared being alone?» and «Who of the boys worked late again?» one must correctly resolve coreference.

Another example is from the field of machine translation:

(4a) We couldn't push a piano$_1$ into the room. It$_1$ was too big.

(4b) We couldn't push a piano into the room$_1$. It$_1$ was too small.

Due to the differentiation of the grammatical gender for inanimate nouns in Russian, the translations differ:

(4a') Оно было слишком большим (Ono bylo slishkom bol'shim)

(4b') Она была слишком маленькой (Ona byla slishkom mal'en'koi)

Choosing the right gender of the pronoun is impossible without coreference resolution.

Currently, the most state-of-the-art NLP algorithms require annotated corpora for training. For coreference resolution, coreference chains annotations are necessary for most of the approaches.

At the moment, we have annotated 523 texts. In the future, we plan to annotate 3000 texts more. This will open more opportunities for using this corpus for training various complex models.

Coreference has several types (in the following examples coreferent mentions are underlined):

- Anaphora

(5) To cook a turkey, you should first rub it all over with the butter.

- Corefering noun phrases

(6) Some of our colleagues are going to be supportive. These kinds of people will earn our gratitude.

- Corefering verb phrases

(7) It is hard to change a human nature. It requires a lot of will.

- Split antecedents

(8) Bob and John are good friends. They go to school together.*

In this paper and in the corpus we present we focus on corefering noun and pronoun phrases like in examples (1-6). We do not cover corefering verb phrases and we do not consider split antecedents as coreferent. Consider the following example.

(9) Peter$_1$ and John$_2$ are classmates. Every day the boys$_3$ are studying together. They$_3$ have lunch and do homework after classes. Peter$_1$ likes geometry and John$_2$ is a fan of poetry.

2. *Related corpora*

2.1. Message Understanding Conference

The first evaluation campaign for the coreference resolution problem was held at the conference Message Understanding Conference-6 in 1995 (for more details see [Grishman, Sundheim 1996]). Participants were to extract named entities from news reports in English and to determine their types (Person, Location, Organization, Time or Quantity). Also they introduced the first coreference resolution quality metric, the MUC score ([Villain et al. 1995]).

2.2. CoNLL-2012 and other notable corpora for English with coreference annotations

Currently, CoNLL-2012 Shared Task [Pradhan et al. 2012] is one of the most well-known evaluation campaigns and a source of training and test data for the coreference resolution task. The corpus used for training and testing contains data in English and Chinese tagged in Universal Dependencies standard and is based on the OntoNotes 5.0 corpus.

At the evaluation event held at the Conference on Computational Natural Language Learning in 2012, the best system in the so-called «closed» track had F-measure around 62 (for English). In the closed track, the systems were required to use only the provided data.

Nowadays, state-of-the-art systems trained on the same data have F-measure 65.7 [Manning, Clark 2016] and 67.7 (see [Lee et al. 2017]). Usually participants use machine learning algorithms (including neural networks algorithms) and perform data pre- and post-processing.

Two other noteworthy corpora for English are OntoNotes and ARRAU

2.3. Other languages

In the last decade similar work was conducted for a plethora of other languages including Basque ([Soraluze et al. 2015]), Polish ([Ogrodniczuk et al. 2013]), and Czech. ([Nedoluzhko 2016]). Since 2016, the annual workshop is being held, CORBON: The Coreference beyond OntoNotes, that explores the state-of-the-art for the coreference resolution for languages other than English (e.g. [Grishina 2017], [Poesio et al. 2018]).

2.4. RuCor — a Russian corpus with coreference annotation

In 2014, the conference "Dialogue" held the coreference resolution competition (for more details see [Toldova et al. 2014]). Participants were provided with a corpus with morphological and coreference annotations consisting of 181 documents. Morphological annotations were done according to the MULTEXT-EAST annotation scheme. The corpus is freely available[1] and is used for various experiments on anaphora and coreference in Russian (e.g. [Khadzhiiskaia and Sysoev 2017], [Toldova and Ionov 2017]).

3. Participants and data sets

3.1. Texts

Texts for the new corpus are taken from the Open Corpus of Russian Language (OpenCorpora)[2] which makes all texts publicly available. The corpus consists of 3769 texts of various genres: fiction, publicist texts, scientific texts, etc. Sampling was performed randomly, therefore, the annotated subcorpus has the similar distribution in genres.

Table 1. Genre distribution of the corpus

chaskor.ru (articles)	8%
chaskor.ru (news)	28%
Wikipedia	10%
Wikinews	15%
Blogs	20%
Fiction	3%
Non-fiction	4%
Legal texts	11%
Other	1%

[1] http://rucoref.maimbava.net/

[2] http://opencorpora.org/

3.2. Coreference corpus description

3.2.1. Layers

The corpus has several layers of annotation:

1. **Coreference chains layer.** For each mention included in a chain with length more than one, there is a line describing it in the following format:

Mention ID→Mention Offset→Mention Length→Chain Id

2. **Mentions layer.** For *each* mention in a text, there is a line describing it in the following format:

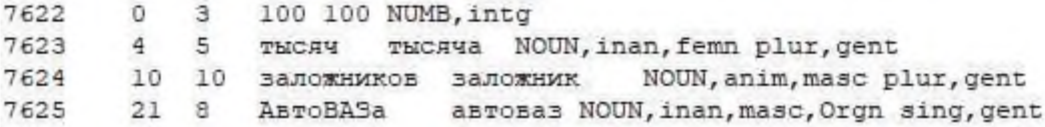
Mention ID→Mention Offset→Mention Length

Mention IDs are sorted in order of appearance in the text.

3. **Morphological layer.** Contains morphological markup from OpenCorpora. For each token, there is a line describing it in the following format:

Token ID→Offset→Length→Token→Lemma→Morph Tags

In the Morph Tags column there are two groups: lexeme features and wordform features. The groups are divided by a space. Tags inside these groups are separated with a comma. The full list of tags can be found on the OpenCorpora website[3]. The annotation is serialised in a CoNLL-like format. Note that punctuation marks are treated as separate tokens; sentences are separated with an empty line.



```
7622    0   3   100 100 NUMB,intg
7623    4   5   тысяч   тысяча  NOUN,inan,femn plur,gent
7624    10  10  заложников  заложник    NOUN,anim,masc plur,gent
7625    21  8   АвтоВАЗа    автоваз NOUN,inan,masc,Orgn sing,gent
```

Fig. 1. Example of morphological markup.

4. **Semantic-syntactic layer.** Automatic tools provided in the ABBYY Compreno suite (for more details see [Anisimovich et al. 2012]) were used to create semantic and syntactic annotations. For each token, there is a line describing it in the following format:

Offset→Token text→Parent Offset→Lemma→Lexical class→Semantic Class→Surface Slot→Semantic Slot→Syntactic Paradigm.

If a token doesn't have a parent (e.g. a predicate of a sentence) then the line does not have "ParentOffset" key. For each semantic class identifier, embeddings are provided. They can be used as features.

[3] http://opencorpora.org/dict.php?act=gram

Mention IDs in the first two layers are the same for the same mentions. At the same time, ABBYY Compreno tokenization may differ from OpenCorpora tokenization. Hence further alignment may be needed for harmonising these two tokenizations in order to use that layer.

E.g. for composite words, ABBYY Compreno tokenization may have several tokens. E.g. *kinoroman* (romance movie) consists of two tokens: *kino* and *roman*.

3.2.2. Coreference annotation

ABBYY Compreno was also used to generate a preliminary automatic markup for the coreference layers (coreference chains and layers).

To generate mention candidates, we used existing tools to perform named entity recognition (see [Stepanova et al. 2016]) and additionally extracted noun and pronoun phrases. Automatic annotations of each document were verified by at least two annotators who could accept a candidate as a mention, deny it or correct its boundaries. Then these two annotated versions of each document were aligned by an assessor.

To generate candidates of coreference chains, we used the output of ABBYY Compreno. This system showed the best results at the previous evaluation event in 2014 (see [Toldova et al. 2014], [Bogdanov et al. 2014]). This preliminary markup was manually checked and corrected.

Average markup coherence measure is presented in Table 2. It was calculated as an F-measure calculated from the comparison of two aligned annotations.

Table 2. Markup coherence measure.

Verifier 1 vs Verifier 2 (mentions)	
coherence	62.7%
Verifier 1,2 vs Final	
coherence	75.5%
lost mentions average (retrieved by another verifier and assessor)	28.1%
lost mentions combined (retrieved by assessor)	8.2%
Verifier 1,2 vs Auto	
Added new mentions average	3.7%
Added new mentions combined	4.3%
Final (mentions) vs Auto	
coherence	47.1%
added new mentions in total	10%
Verifier 1,2 + Assessor + Chains (mentions) vs Auto	
added new mentions in total	12.9%

The reason behind low coherence measure of Final (mentions) vs Auto might be due to low precision of the automatic markup in comparison to the final markup. It showed precision of 31.9% (with 90% recall). We might need to improve the precision of automatic markup or increase the number of verifiers of each document.

There was only one verifier tagging chains based on the final version of the first layer of the markup and the automatic coreferent chains markup. Hence there are no coherence measure for chains.

3.2.3. Basic principles and instructions of gold standard preparation

In the corpus, we considered as mentions the following:

- Named entities: persons, organizations, locations
- Common nouns and noun phrases
- Pronouns, except negative pronouns and the reflexive pronoun *seb'ya* (*oneself*)

Note that the list lacks facts and verb phrases. In some corpora, the coreference annotation includes these, but in this corpus we decided not to annotate them.

Also note that we do not distinguish mention types in our annotation scheme, they all are annotated as mentions. At the same time, annotation strategies differ.

In case of named entities, we include:

- Names, e.g. *Peter, Everest, Google*
- noun phrases composed of a named entity and its specifiers, e.g. *internet-gigant* Gugl (*internet giant* Google), *mal'chik* Pet'ya (the *boy* Peter), *gora* Ever'est (the *mountain* Everest). We don't add terms which refer to another mention, e.g. *ot'ets** P'eti (Peter's *father**)

In the case of common nouns and noun phrases, we use the full noun phrase, e.g. *krasivyi mal'chik, kotoryi prodayot knigi* (a *pretty boy selling books*)

We decided not to annotate reflexive (*oneself*) and reciprocal (*each other*) pronouns and we do not include them in coreference chains. Also we don't unite mentions if one of them refers to the entity that includes as a part the entity another mention refers to.

See example 9 for demonstration.

(9) Peter$_1$ and John$_2$ are classmates. Every day the boys$_3$ are studying together. They$_3$ have lunch and do homework after classes. Peter$_1$ likes geometry and John$_2$ is a fan of poetry.

3.3. Anaphora corpus description

For the anaphora resolution track we used the same corpus but with only anaphoric links. As such we considered only pairs between mentions in the same chain where the second element is a 3rd gender pronoun.

Since there was no original anaphoric annotation, and it was impossible to extract the exact pairs, we used parts of coreference chains from the first item in a chain until the last pronoun there.

3.4. Train and test sets

Currently, the train set contains 395 random documents, Test set – 127 random documents. They are available for downloading from a Google Drive folder[4].

Table 3. Train set statistics.

Mentions	21486
Mentions in coreference chains	18282
Coreference chains	4110

Table 4. Test set statistics.

Mentions	7475
Mentions in coreference chains	6877
Coreference chains	1568

4. Evaluation metrics

4.1. Metrics for coreference resolution MUC, B-Cubed, CEAF-E

We use three measures for the coreference track evaluation: MUC [Vilain et al, 1995], B^3 [Bagga, Baldwin, 1998] and CEAF-e [Luo, 2005]. In [Moosavi, Strube 2016] the overview of these metrics is presented.

First, we check consistency of results. We check:

- how many numbers in each row (must be 4),
- if mentions repeat in one document,
- if there are chains with a single mention.

If there are inconsistencies in the results, the script outputs a warning with a file name and excludes the file.

[4] Coreference Corpus w/o test answers
https://drive.google.com/open?id=1xThOmNy_o1Vtw4dlr6kyHhM9hFdw_nb7
Test answers https://drive.google.com/open?id=1jqEeIS39pKtbIF9jh8T27YJEXINSH7EH

If there is a division by zero during the procedure metrics are considered to be equal to zero.

If both recall and precision equal to zero, then we consider F-measure equal to zero as well.

If some chain ID is omitted in response (e.g. 1, 2, 4, 5) the script outputs a warning with a file name. Metrics for such documents will be calculated incorrectly.

The evaluation script can be downloaded from a Google Drive folder[5].

Then we align mentions from gold standard and results. Two mentions are aligned if and only if their offsets and lengths coincide. Unaligned mentions are considered missed or spurious.

Finally, metrics are calculated.

4.2. Measures for anaphora resolution

As was mentioned before, for the anaphora resolution track, only the third person pronouns from the test set are considered in the scoring procedure (there were occasionally missed pronouns in the test data set). Since there was no separate anaphoric annotations and it is impossible to know, which element of a coreference chain is the antecedent, we use two types of scoring for anaphora resolution. Any previous mention from the chain before a particular pronoun is considered to be correct antecedent for soft scoring, and only the closest previous mention for the strict scoring. In case of unaligned mentions as the antecedent of a pronoun the answer of a system is treated as false positive. The case of omittance of a pronoun by a system is considered to be false negative. As a consequence, the accuracy (the number of true answers of a system / all the 3rd person pronouns, annotated in the gold standard set) coincides with the recall.

5. Results

5.1. Coreference track results

Six teams participated in Coreference track:

- legacy
- SagTeam
- DP
- Julia Serebrennikova
- MorphoBabushka

Table 5. Coreference track results.

Team	muc	bcube	ceafe	mean
legacy	75.83	66.16	64.84	68.94

[5] https://drive.google.com/open?id=1w6406JU78taPfL28ZmpyE3JdVvWN60L3

SagTeam	62.23	52.79	52.29	55.77
DP	62.06	53.54	51.46	55.68
DP (additionally trained on RuCor)	**82.62**	**73.95**	**72.14**	**76.24**
Julia Serebrennikova	48.07	34.7	38.48	40.42
MorphoBabushka	61.36	53.39	51.95	55.57

5.2. Anaphora track results

Some of the teams submit the results only for the anaphora track, some others submit the results for both tracks. Moreover, there were teams that submitted the results for several runs under different conditions. The micro-averaged results are presented in the Table 6.

Table 5. Anaphora track results.

Team	Run	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
		soft				strong			
DP	Full	76.30%	79.20%	76.30%	77.80%	68.10%	70.70%	68.10%	69.40%
	On gold	**91.00%**	**91.40%**	**91.00%**	**91.20%**	**83.50%**	**83.90%**	**83.50%**	**83.70%**
Etap		52.40%	78.70%	52.40%	62.90%	39.10%	58.70%	39.10%	46.90%
Legacy		70.80%	75.70%	70.80%	73.20%	59.10%	63.10%	59.10%	61.00%
NSU_ai		23.20%	43.30%	23.20%	30.20%	6.90%	12.90%	6.90%	9.00%
Morphobabushka	best-muc-1	62.90%	63.50%	62.90%	63.20%	38.80%	39.10%	38.80%	39.00%
	best_b3f1_and_ceafe_4	55.10%	57.30%	55.10%	56.20%	37.10%	38.60%	37.10%	37.80%
	best_b3f1_and_ceafe_5	54.50%	59.40%	54.50%	56.80%	35.10%	38.30%	35.10%	36.60%
Meanotek		44.40%	58.70%	44.40%	50.60%	34.70%	45.80%	34.70%	39.40%
		52.40%	78.70%	52.40%	62.90%	39.20%	58.80%	39.20%	47.00%

As compared to the results of Ru-eval 2014 the best F-measure score is higher (83.7% vs. 76%), although given completely different settings, these results cannot be compared directly.

6. *Conclusions*

In this paper, we have introduced a new open Russian coreference resolution dataset that was used for training and evaluation of participants systems. We have outlined two tasks and evaluation metrics. We have described the design of the dataset and provided means for metrics computation. The size of the corpus (5.7k chains with 25k mentions) allows to successfully use machine learning methods for both mention detection and coreference or anaphora resolution.

Since RU-EVAL 2014, the NLP technologies and resources for Russian has changed greatly. We have estimated the current state-of-the-art for two mentioned NLP tasks (83.7% for Anaphora and 76.24% for Coreference resolution).

We hope that RU-EVAL 2019 has provided necessary means for further progress in the field and has helped the participants to achieve better results and deeper understanding of the problem.

7. Acknowlegments

We would like to thank Irina Sokolova, Daria Marinina, and Julia Serebrennikova who also participated in corpus preparation. We also are grateful to Vladimir Selegei, Svetlana Mazurina, and Anna Nedoluzhko for their help in the event discussions and organization. We are also most grateful to the participants of the forum. The work was partially funded by ABBYY.

References

Anisimovich K., Druzhkin K., Minlos F., Petrova M., Selegey V., and Zuev K. (2012), Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference ”Dialogue”, vol. 11, pp. 91– 103.

Bagga, Baldwin 1998 — Bagga, A., & Baldwin, B. (1998, August). Entity-based cross-document coreferencing using the vector space model. In Proceedings of the 17th international conference on Computational linguistics-Volume 1(pp. 79-85). Association for Computational Linguistics.

Bogdanov et al. 2014 — Bogdanov, A., Dzhumaev, S., Skorinkin, D., & Starostin, A. (2014). Anaphora analysis based on ABBYY Compreno linguistic technologies. Computational Linguistics and Intellectual Technologies, 13(20), 89-101.

Cai, Strube 2010 — Cai, J., & Strube, M. (2010, September). Evaluation metrics for end-to-end coreference resolution systems. In Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue (pp. 28-36). Association for Computational Linguistics.

Grishina, Y., 2017. CORBON 2017 Shared Task: Projection-Based Coreference Resolution. In Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017) (pp. 51-55).

Grishman, Sundheim 1996 — Grishman, R., & Sundheim, B. (1996). Message understanding conference-6: A brief history. In COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics (Vol. 1).

Khadzhiiskaia, A., Sysoev, A. (2017). Coreference resolution for Russian: taking stock and moving forward. In 2017 Ivannikov ISPRAS Open Conference (ISPRAS), pp. 70-75. IEEE, 2017.

Lee et al 2017 — Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). End-to-end Neural Coreference Resolution. arXiv preprint arXiv:1707.07045.

Luo 2005 — Luo, X. (2005, October). On coreference resolution performance metrics. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (pp. 25-32). Association for Computational Linguistics.

Manning, Clark 2015 — Clark, K., & Manning, C. D. (2015, July). Entity-Centric Coreference Resolution with Model Stacking. In ACL (1) (pp. 1405-1415).

Martschat, Strube 2015 — Martschat, S., & Strube, M. (2015). Latent structures for coreference resolution. Transactions of the Association for Computational Linguistics, 3, 405-418.

Moosavi, Strube 2016 — Moosavi, N. S., & Strube, M. (2016). Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric. In ACL (1)

Nedoluzhko 2016 — Ngụy Giang Linh, Michal Novak, Anna Nedoluzhko (2016). Coreference Resolution in the Prague Dependency Treebank. (ÚFAL/CKL Technical Report #TR-2011-43). Prague: Universitas Carolina Pragensis.

Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A. and Zawisławska, M., 2013, December. Polish coreference corpus. In Language and Technology Conference (pp. 215-226). Springer, Cham.

Poesio, M., Ng, V. and Ogrodniczuk, M., 2018. Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference. In Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference.

Pradhan et al. 2012 — Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., & Zhang, Y. (2012, July). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In

Joint Conference on EMNLP and CoNLL-Shared Task (pp. 1-40). Association for Computational Linguistics.

Soraluze, A., Arregi, O., Arregi, X. and de Ilarraza, A.D., 2015. Coreference Resolution for Morphologically Rich Languages. Adaptation of the Stanford System to Basque. Procesamiento del Lenguaje Natural, 55, pp.23-30.

Stepanova M. E., Budnikov E. A., Chelombeeva A. N., Matavina P. V., Skorinkin D. A. (2016),Information Extraction Based on Deep Syntactic-Semantic Analysis. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference ”Dialogue”, pp. 721-732.

Toldova et al. 2014 — Toldova, S., Roytberg, A., Ladygina, A., Vasilyeva, M., Azerkovich, I., Kurzukov, M., ... & Grishina, Y. (2014). RU-EVAL-2014: Evaluating anaphora and coreference resolution for Russian. Computational Linguistics and Intellectual Technologies, 13(20), 681-694.

Toldova, S. and Ionov, M., 2017. Coreference resolution for russian: the impact of semantic features. In Proceedings of International Conference Dialogue-2017 (pp. 348-357).

Vilain et al, 1995 — Vilain, M., Burger, J., Aberdeen, J., Connolly, D., & Hirschman, L. (1995, November). A model-theoretic coreference scoring scheme. In Proceedings of the 6th conference on Message understanding (pp. 45-52). Association for Computational Linguistics.